DOCUMENT RESUME

ED 467 378 TM 034 304

AUTHOR	van der Linden, Wim J.; Vos, Hans J.; Chang, Lei
TITLE	Detecting Intrajudge Inconsistency in Standard Setting Using Test Items with a Selected-Response Format. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-02-01
PUB DATE	2000-00-00
NOTE	24p.; Some data were provided by the National Institute of Educational Measurement.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Interrater Reliability; *Judges; *Probability; *Standard Setting; Test Format; *Test Items

ABSTRACT

In judgmental standard setting experiments, it may be difficult to specify subjective probabilities that adequately take the properties of the items into account. As a result, these probabilities are not consistent with each other in the sense that they do not refer to the same borderline level of performance. Methods to check standard setting data for intrajudge inconsistencies are thus of paramount importance to setting meaningful standards. This paper presents a method of consistency analysis for standard setting experiments in which judges specify probabilities for each response alternative of the items. The method is based on a residual diagnosis of the subjective probabilities under the hypothesis of a consistent judge to the probabilities. An empirical example shows how the method can be used to identify sources of inconsistency in response alternatives, items, or judges. (Contains 19 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 378

TM

Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format

Research Report 02-01

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Wim J. van der Linden
Hans J. Vos

Lei Chang
Chinese University of Hong Kong

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034304

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format

Wim J. van der Linden

Hans J. Vos

Lei Chang

Chinese University of Hong Kong

The authors are indebted to the National Institute of Educational Measurement (Citogroep), Arnhem, The Netherlands, for making available the data set in the empirical example and to Wim M.M. Tielen for his computational support.

Abstract

Intrajudge Inconsistency in Standard Setting

In judgmental standard setting experiments, it may be difficult to specify subjective probabilities that adequately take to the properties of the items into account. As a result, these probabilities are not consistent with each other in the sense that they do not refer to the same borderline level of performance. Methods to check standard setting data for intrajudge inconsistencies are thus of paramount importance to setting meaningful standards. This paper presents a method of consistency analysis for standard setting experiments in which judges specify probabilities for each response alternatives of the items. The method is based on a residual diagnosis of the subjective probabilities under the hypothesis of a consistent judge to the probabilities. An empirical example shows how the method can be used to identify sources of inconsistency in response alternatives, items, or judges.

Keywords: Angoff Method; Interdependent Evaluation of Alternatives; Intrajudge Inconsistency; Polytomous Response Models; Nedelsky Method; Standard Setting.

Introduction

The question of how to justify standards for educational tests or assessments has been a persistent source of debate among educators and measurement specialist. The dominant view since a discussion in a 1978 special issue of the *Journal of Educational Measurement* (Glass, 1978; Hambleton, 1978; Popham, 1978) is that standards can not be justified by an independent criterion but that their justification should follow from an evaluation of the procedure used to set them. Though the question of what constitutes a good standard setting method still is controversial, it seems safe to assume that for judgmental standard setting methods the requirement of the judgments being consistent with the objective properties of the test items has universal validity.

Several types of inconsistent judgemental behavior in standard setting are possible (van der Linden, 1996). For example, in an experiment in which judges evaluate actual test booklets of examinees, they may specify that Booklet A demonstrates more proficiency than Booklet B, Booklet B more than Booklet C, but Booklet C less than Booklet A. Likewise, in an Angoff (1971) experiment, a judge may specify higher probabilities of success for items that are more difficult. This type of inconsistency parallels the one in an Nedelsky (1954) experiment with a judge eliminating more options for a more difficult item. Inconsistency may also happen in standard setting experiments with two tests or assessment instruments. If the standard set on one instrument can not be predicted from the one on the other by a (possibly nonlinear) regression equation that excellently fits the bivariate distribution of response data for both instruments, these standards are inconsistent.

Each of these examples points to inconsistent behavior in standard setting that can be characterized as *intrajudge* inconsistency. Besides, the term *interjudge* inconsistency has been used to describe differences in standards between judges in the same experiment. Though analyses of interjudge consistency may be useful in standard setting experiments where judges are provided with meticulously defined performance levels, the requirement that different judges should set the same standard does not have the universal validity the requirement of intrajudge consistency has.

It is the purpose of this paper to introduce a method for analyzing intrajudge inconsistencies in standard setting experiments where the judges are required to specify

probabilities of an examinee functioning at the borderline level of performance for each response alternative of the items. The standard setting method used in the empirical example was the method of interdependent evaluation of all response alternatives (IDEA) (Chang, van der Linden & Vos, 2001). The method combines positive and avoids negative features of the Angoff and the Nedelsky method. Like the Nedelsky method, it forces the judges to look into all alternatives and to evaluate the behavior of borderline examinees with respect to each of them. On the other hand, like the Angoff method, it allows judges to specify probabilities on the full scale from [0,1] and avoids the problems of discreteness inherent in the Nedelsky method. However, the standard is calculated only from the probabilities for the correct alternative. Though not highlighted in this paper, the proposed method of analyzing intrajudge inconsistency can also be used in standard setting experiments with polytomously scores items. The only difference would be a possible other choice for the item response theory (IRT) model used in the empirical example below.

The method is based on the technique of residual analysis in statistics. It requires a model for the probabilities on the alternatives be fit to response data from a representative set of examinees and then analyzes the residual probabilities under the hypothesis of consistent judgements. This type of analysis was introduced in van der Linden (1982; see also Kane, 1987) for standard setting experiments based on the Angoff or Nedelsky method. The current paper generalizes the applicability of the analysis to standard setting experiments that exploit the full set of response alternatives. One of the advantages of this generalization is that it is now possible not only to identify sources of inconsistency that reside in the judge or in the items but also in specific response alternatives or in interactions between judges and specific alternatives. In fact, a surprising finding in the empirical example in this paper is that nearly all judges had systematically greater difficulty dealing with the correct than with the incorrect alternatives of the items.

Definitions and Notation

The test items used in the standard setting experiment are denoted as $i = 1, ..., n$, with the response alternatives for item i denoted as $k_i = 1, ..., m_i$. A separate notation is needed for the correct and wrong alternatives of the items. The correct alternative of item

i is denoted as g_i, while an arbitrary incorrect alternative is denoted as w_i. The items are assumed to measure a (unidimensional) variable θ representing the performances of the examinees. Each of the judges $j = 1, ..., N$ is asked to choose a standard for the performance level required from the examinees. The standard for judge j is denoted as a cut-off score θ_{cj}. Observe that the standards are indexed by j because different judges may have different standards. If the standard setting experiment is required to result in a single standard for the whole panel of judges, some form of consensus making has to be introduced in the standard setting process that results in a common choice $\theta_{cj} = \theta_c$ for all judges. Alternatively, a statistical operation, for instance, averaging of individual cutoff scores, can be used to combine all individual standards into a single standard.

For each item the judges are required to specify the probabilities of an examinee operating at performance level θ_{cj} to produce response X_i=k_i on item i. Because this probability is the result of a judgmental process, it is denoted as $p^s_{k_i j}$, with superscript s to indicate its subjective nature. The fact that these probabilities are required to sum to 1 forces the judges to coordinate their specifications between the alternatives.

IRT Model

If the response data for the populations of examinees fit an IRT model for items, with a polytomous response format, we also have objective probabilities for response X_i=k_i by an examinee at performance level θ_{cj}. In the empirical example below, Thissen and Steinberg's model for multiple-choice items (Thissen & Steinberg, 1984, 1997) was fitted to the data. The model defines the probability of an examinee at θ_{cj} producing response $X_i = k_i$ as:

$$p_{k_i j} \equiv \Pr\{X_i = k_i \mid \theta_{cj}\} \equiv \frac{\exp\{a_{k_i}(\theta_{cj} - b_{k_i})\} + d_{k_i}\exp\{a_{0_i}(\theta_{cj} - b_{0_i})\}}{\sum_{h_i=0}^{m_i}\exp\{a_{h_i}(\theta_{cj} - b_{h_i})\}}, \quad (1)$$

where b_{k_i} and a_{k_i} are the location and discriminating power of alternative k of item i, respectively. The model, which generalizes Bock's (1997) nominal response model, was chosen because of its flexibility to deal with guessing on multiple-choice items. It does so by assuming that among the examinees that give response k_i to item i an (a priori unknown) proportion d_{k_i} guesses ($\sum_{k_i=1}^{m_i} d_{k_i} = 1$). The process of guessing is not assumed to be blind but to be dependent on θ with probabilities given by $\exp\{a_{0_i}(\theta_{cj} -$

$b_{0_i}\}/\sum_{h_i=0}^{m_i}\exp\{a_{h_i}(\theta_{cj}-b_{h_i})\}$, with a_{0_i} and b_{0_i} denoting the location and discriminating power of the response function for the examinees who guess.

When the model in (1) is fitted to data from achievement tests, the response function for the correct alternative should be monotone in θ. In the application below, the validity of this assumption is tested against the alternative of a nonmonotone response function.

Error Definition

Observe that θ_{cj} should be calculated from the subjective probabilities provided by judge j *under the hypothesis of consistent judgments*. The assumption is typical of the technique of residual analysis used in this paper. The steps in this technique are: First, a model for the probabilities on the alternatives is fitted to the response data from a representative set of examinees. In the application below, the model is the one specified in (1). Second, under the null hypothesis of a consistent judge a cutoff score is fitted to his/her subjective probabilities of the judge. Third, the residuals, that is, the differences between the objective probabilities from the model and the subjective probabilities form the judge, are calculated. Fourth, the residuals are analyzed for inconsistencies, and potential explanations of the inconsistencies are developed.

For the current response model in (1), the calculation of the cutoff score θ_{cj} for judge j in the second step is based on the following operations:

1. Summing the probabilities $p^s_{g_i j}$ over the items in the test;

2. Summing the objective probabilities for the correct alternatives over the items in the test;

3. Equating the two sums and calculating θ_{cj} as the root of the equation.

That is, θ_{cj} is calculated as the root of:

$$\sum_{i=1}^{n} p^s_{g_i j} = \sum_{i=1}^{n} \frac{\exp\{a_{g_i}(\theta_{cj}-b_{g_i})\} + d_{k_i}\exp\{a_{0_i}(\theta_{cj}-b_{0_i})\}}{\sum_{k_i=0}^{m_i}\exp\{a_{k_i}(\theta_{cj}-b_{k_i})\}}. \tag{2}$$

The error by judge j on alternative k of item i is thus equal to the residual probability

$$e_{k_i j} \equiv p^s_{k_i j} - p_{k_i j}. \tag{3}$$

It is now possible to aggregate the error in (3) over response alternatives, items, or judges. This aggregation results in inconsistency indices for (combinations of) judges and items. We first introduce a set of unstandardized inconsistency indices and then indicate how to standardize these indices to take possible values only in the interval [0,1].

Errors by Individual Judges

The absolute errors by judge j on the correct and incorrect alternatives of item i are given by

$$\epsilon_{gij} \equiv \left| p^s_{gij} - p_{gij} \right| \tag{4}$$

and

$$\epsilon_{wij} \equiv (m_i - 1)^{-1} \sum_{k=1;k\neq g}^{m_i} \left| p^s_{kij} - p_{kij} \right|, \tag{5}$$

respectively.

Aggregating these errors over the items gives the following indices for the average errors by judge j on the correct, incorrect and across all alternatives at the level of the test:

$$\epsilon_{gj} \equiv n^{-1} \sum_{i=1}^{n} \left| p^s_{gij} - p_{gij} \right| \tag{6}$$

$$\epsilon_{wj} \equiv (\sum_{i=1}^{n} m_i - n)^{-1} \sum_{i=1}^{n} \sum_{k=1;k\neq g}^{m_i} \left| p^s_{kij} - p_{kij} \right| \tag{7}$$

$$\epsilon_{j} \equiv (\sum_{i=1}^{n} m_i)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_i} \left| p^s_{kij} - p_{kij} \right| \tag{8}$$

This choice for absolute values of the errors is made to prevent them from compensating each other when they are aggregated within or between items or judges.

Errors by Panel of Judges

Test items differ in the likelihood of a judge making an error in his/her specification of the subjective probabilities. The reason for such differences may reside, for example, in sloppy behavior by the judge, but also in the formulation of the items, the difficulty of the correct alternative, or the familiarity of the judge with specific topics in the domain tested. Item analysis based on errors aggregated over the panel of judges can help to reveal the actual sources of such differences.

The following equations give the average errors on the correct alternative, incorrect alternatives, and across all alternatives of item i across the panel of judges:

$$\epsilon_{g_i} \equiv N^{-1} \sum_{j=1}^{N} \left| p^s_{g_i j} - p_{g_i j} \right| \tag{9}$$

$$\epsilon_{w_i} \equiv N^{-1}(m_i - 1)^{-1} \sum_{j=1}^{N} \sum_{k=1;k\neq g}^{m_i} \left| p^s_{k_i j} - p_{k_i j} \right| \tag{10}$$

$$\epsilon_i \equiv (N m_i)^{-1} \sum_{j=1}^{N} \sum_{k=1}^{m_i} \left| p^s_{k_i j} - p_{k_i j} \right| \tag{11}$$

Analogous to (6)-(8), the errors by a panel of judges can be aggregated over all items in the test. These aggregates can be used, for example, to detect differences between the error levels for the correct and incorrect alternatives or the general error level for the panel of judges on the test. The equations are:

$$\epsilon_g \equiv (Nn)^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} \left| p^s_{g_i j} - p_{g_i j} \right| \tag{12}$$

$$\epsilon_w \equiv (N)^{-1} \left(\sum_{i=1}^{n} m_i - n \right)^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} \sum_{k=1;k\neq g}^{m_i} \left| p^s_{k_i j} - p_{k_i j} \right| \tag{13}$$

$$\epsilon \equiv (N\sum_{i=1}^{n} m_i)^{-1} \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{k=1}^{m_i} |p^s_{k_i j} - p_{k_i j}| \tag{14}$$

Standardized Consistency Indices

The above inconsistency indices should be used descriptively. The problem of how to specify a statistical test for the hypothesis of consistent judgements is still hampered by the fact that it appears to be difficult to formulate a valid statistical model for the distribution of the subjective probabilities $p^s_{k_i j}$ across replications. To support the comparison of errors between items, judges, or occasions, it is therefore important to have standardized versions of the indices that have a common range of possible values.

Standardization of the above indices to indices that can take values in the full interval [0,1] is achieved through the following transformation

$$C \equiv \frac{M_\epsilon - \epsilon}{M_\epsilon}, \tag{15}$$

where ϵ is a generic symbol for the inconsistency indices and M_ϵ is the maximum value of the index possible. The maximum is found if index ϵ is calculated with the expression $p^s_{k_i j} - p_{k_i j}$ in (3) replaced by

$$\max\{p_{k_i j}, 1 - p_{k_i j}\}. \tag{16}$$

Because the calculations are straightforward, no equations for the consistency indices are given.

The main purpose for standardizing the residuals is to make them independent of the objective probabilities of success at the performance level of the borderline examinee, θ_{cj}. The maximum residual in (16) varies as a function of θ, whereas index C does not. Observe that the direction of C is also opposite to the direction of ϵ. C should therefore be considered as a consistency index; the closer its value to 1, the more consistent the judgments. The maximum C=1 is obtained if at θ_{cj} it holds that $p^s_{k_i j} = p_{k_i j}$ for all alternatives, items, and/or judges over which the index is defined.

Empirical Example

A standard setting experiment was conducted in which eight judges used the method of interdependent evaluation of items alternatives (IDEA) to set a pass-fail standard on a test of German as a second language consisting of items used previously in a national school leaving exam at the end of secondary education in the Netherlands. The purpose of this small experiment was not to set an actual standard for the exam or to assess the typical error level of judges operating in a standard setting experiment, but only to illustrate the type of residual analysis of intrajudge inconsistency advocated in this paper.

The original test used in the national exam had 29 items with 3-5 response alternatives. The items were calibrated under the model in (1) using the response data from 16,1648 examinees and the software program *Multilog* (Thissen, 1991). The goodness of fit of the model was assessed both against a less restrictive and more restrictive model fitted on the same data set. The direct likelihood-ratio test of the model against the general multinomial alternative in *Multilog* could not be used because the number of examinees was of much smaller order than the number of possible response patterns (1.938×10^{11}). For the use of such alternative goodness-of-fit tests, see Thissen and Steinberg (1984). The less restrictive model was Mokken's (1997) nonparametric response model. This model was used to check the items for unidimensionality of θ as well as monotonicity of the response function for the correct alternative using the software program *MSP 5* (Molenaar & Sijtsma, 2000). A set of 19 items yielded a scalability coefficient H=.14, which is to be considered as a conservative value (Molenaar & Sijtsma, 2000). Because the Mokken model does not assume any parametric form for the response functions, it follows that the data support these two critical assumptions. The assumption of monotonicity of the response functions for the correct alternatives is particularly important because the model in (1) was applied to achievement test items. The more restrictive model was the nominal response model (Bock, 1997). For the same set of items, a likelihood-ratio test showed that this model had to be rejected in favor of the model in (1) ($p < .001$). This set was therefore used in the experiment.

The judges were secondary school teachers with an average of 8.5 years of experience in teaching German. The judges were trained using realistic exercises until each of them declared to be competent in the task. It is believed that both the selection and training of

the judges qualifies them for the standard setting experiment (Raymond & Reid, 2001).

[Table 1 and 2 about here]

Summaries of the (aggregated) residuals for the correct and incorrect alternatives are given in Table 1 and 2, respectively. A consistent trend in the two tables is the difference between the residuals for the correct and incorrect alternatives. The average residuals across all judges and items is .18 for the correct and .11 for the incorrect alternatives. The ranges for the average residuals per judge are remarkably small: (.16-.23) for the correct alternatives and (.10-.13) for the incorrect alternatives. The difference in range can be explained by the fact that the results for the incorrect alternative are based on an extra step of averaging (the number of incorrect alternatives was 2-4).

The average residuals per item are also in a relatively small range for the incorrect alternatives, (.07-.18). However, the average residuals per item for the correct alternatives showed two outlying results: .35 for Item 2 and .45 for Item 12. If these results are neglected, the range runs from (.07-.22).

A comparison between the residuals for Item 2 and 12 in Tables 1-2 shows that both are uniformly high across judges for the correct alternative. Item 12 also shows uniformly high residuals for the incorrect alternative, whereas Item 2 shows results for the judges that are not systematically larger than from those for the other items. There are two reasons why residuals can be large: (1) attributes specific to the item that make it difficult to specify subjective probabilities for one or more of its alternatives; and (2) the dependency of the residuals on θ_{cj}.

[Table 3 and 4 about here]

The latter explanation can be rejected if the analysis is based on standardized consistency indices. Table 3 and 4 shows the values of these indices for the same items and judges. A comparison between these two sets of tables seems to support the hypothesis that the results for Item 12 are due to the attributes of the item, in particular, attributes of the correct alternatives (the values for the incorrect alternative do not show any remarkable pattern). The results for Item 2 where more in line with those for the others items (albeit that the average consistency across judges is among the lowest values). Getting back to the response data for the examinees, the authors found that the p-value

for these examinees (.40) was lower than the a-value for one of the incorrect alternatives (.49). This observation may suggest an ambiguity in the correct alternative. In a real-life application of this method, with feedback to the judges, the next step would be to ask the judges to discuss this alternative. If the conclusions did not converge, or if they indicated a technical error in this alternative, the natural decision would be to remove the item form the set and ask the judges to reconsider their subjective probabilities on the other items only.

Concluding Observations

A systematic trend in the results in Table 1-4 is more consistent behavior for the incorrect than for the correct alternatives of the items. This trend seems to hold for nearly each judge (the only clear exception is Judge 5). The fact that this trend holds for the standardized consistency indices as well as for the residuals seems to exclude explanations based on differences in response probabilities for examinees performing at the cutoff scores θ_{cj}. As a tentative explanation, it is suggested that correct alternatives are more difficult to comprehend than incorrect alternatives and that the judges were therefore less capable of specifying probabilities of success on items.

It is not known if this trend generalizes to other content domains. If it would, an interesting practical conclusion would be to set standards using probabilities on the incorrect rather than the correct alternatives. The standard on the θ scale should then be calculated from a version of (2) where in the left- and right-hand side the sums are defined over the most consistent incorrect alternative. Or as an average over a subset of consistent incorrect alternatives. The standard on the number-correct scale cutoff score would then follow from the one on the θ scale via the right-hand side of the current version of (2). This method would amount to a continuous version of the Nedelsky technique. In fact, the method of interdependent evaluation of alternatives used in the empirical example is flexible enough to make a *post hoc* decision on what probabilities to use in the definition of the sums in (2), that is, after all probabilities have been obtained and it is known on which alternative the judges have operated most consistently.

In the empirical example, the objective probabilities $p_{k_i j}$ were calculated using estimates for the parameters in the response model in (1). Though it is possible to calculate

confidence intervals or posterior highest density intervals for the values in Table 1-4 to account for estimation error, this was not done. The number of examinees used to fit the model to the response data was large enough to ignore estimation error. Also, with such intervals, the users may be tempted to interpret the results as if they are from statistical tests, while, as indicated earlier, they should be used only to describe the consistency of judges in the standard setting experiment.

As already alluded to earlier, implementations of standard setting experiments in real life typically have several stages in which judges are encouraged to reconsider their subjective probabilities based on feedback they receive from the facilitator of the process (Reckase, 2001). The proposed use of the residual analysis introduced in this paper is as part of this type of feedback in a multi-stage experiment. It is believed that the format used in Tables 1-4 is easy to understand by the judges typically used in such experiments.

References

Angoff, W. H. (1971). Scales, normes and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.; pp. 508-600). Washington, DC: American Council on Education.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer.

Chang, L., van der Linden, W. J., & Vos, H. J. (2001). *A new test-centered standard setting method based on interdependent evaluation of item alternatives.* Manuscript submitted for publication.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15,* 277-290.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. (1987). On the use of IRT models with judgemental standard setting procedures. *Journal of Educational Measurement, 24,* 333-345.

Mokken, R. L. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP 5 for Windows: A program for Mokken scale analysis for polytomous items.* Groningen, The Netherlands: iecProGAMMA.

Nedelsky, L. (1954). Absolute grading for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15,* 297-300.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 159-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. (1991). *Multilog user's guide*. Chicago, IL: Scientific Software, Inc.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501-519.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

van der Linden, W. J.(1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308. (Corrigenda in *Journal of Educational Measurement, 23,* 265-266.)

van der Linden, W.J. (1996). A conceptual analysis of standard setting in large-scale assessments. In M.L. Bourque (Ed.) *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments* (Vol. 2, pp. 97-118). Washington, DC: National Assessment Governing Board & National Center for Education Statistics.

TABLE 1
Summary of residuals for correct alternative

Item	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Average
1	0.28	0.19	0.34	0.19	0.15	0.22	0.09	0.24	0.21
2	0.24	0.50	0.37	0.43	0.22	0.41	0.35	0.30	0.35
3	0.16	0.06	0.07	0.09	0.21	0.04	0.07	0.01	0.09
4	0.31	0.13	0.01	0.13	0.01	0.40	0.27	0.34	0.20
5	0.08	0.00	0.10	0.11	0.14	0.11	0.09	0.07	0.09
6	0.04	0.13	0.20	0.24	0.02	0.17	0.07	0.02	0.11
7	0.17	0.09	0.13	0.12	0.32	0.35	0.21	0.16	0.19
8	0.18	0.20	0.19	0.14	0.26	0.34	0.15	0.17	0.20
9	0.02	0.05	0.01	0.29	0.12	0.12	0.01	0.09	0.09
10	0.23	0.21	0.13	0.40	0.00	0.20	0.39	0.18	0.22
11	0.06	0.14	0.10	0.18	0.11	0.25	0.18	0.39	0.18
12	0.42	0.41	0.45	0.59	0.37	0.45	0.47	0.45	0.45
13	0.04	0.03	0.03	0.14	0.07	0.21	0.01	0.03	0.07
14	0.11	0.14	0.18	0.32	0.22	0.13	0.06	0.30	0.18
15	0.14	0.08	0.09	0.05	0.09	0.18	0.22	0.22	0.13
16	0.09	0.05	0.03	0.37	0.06	0.01	0.13	0.06	0.10
17	0.18	0.40	0.40	0.02	0.20	0.19	0.12	0.23	0.22
18	0.03	0.02	0.18	0.38	0.26	0.14	0.10	0.15	0.16
19	0.23	0.22	0.00	0.17	0.15	0.24	0.22	0.14	0.17
Average	0.16	0.16	0.16	0.23	0.16	0.22	0.17	0.19	0.18

TABLE 2
Summary of residuals for incorrect alternatives

Item	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Average
1	0.13	0.09	0.12	0.08	0.05	0.11	0.11	0.09	0.09
2	0.11	0.17	0.12	0.14	0.07	0.14	0.12	0.10	0.12
3	0.09	0.08	0.03	0.02	0.10	0.03	0.03	0.06	0.06
4	0.11	0.10	0.08	0.19	0.06	0.20	0.17	0.11	0.13
5	0.10	0.12	0.17	0.04	0.23	0.10	0.08	0.19	0.14
6	0.12	0.09	0.07	0.08	0.10	0.07	0.05	0.04	0.08
7	0.10	0.21	0.07	0.07	0.18	0.12	0.22	0.13	0.17
8	0.11	0.12	0.06	0.13	0.10	0.12	0.06	0.13	0.10
9	0.11	0.06	0.05	0.26	0.06	0.12	0.05	0.04	0.09
10	0.14	0.16	0.09	0.17	0.02	0.10	0.18	0.07	0.12
11	0.21	0.04	0.17	0.18	0.14	0.24	0.09	0.24	0.16
12	0.18	0.14	0.17	0.20	0.22	0.20	0.16	0.22	0.18
13	0.03	0.01	0.06	0.10	0.10	0.07	0.12	0.06	0.07
14	0.10	0.06	0.09	0.13	0.08	0.23	0.05	0.10	0.10
15	0.08	0.09	0.03	0.05	0.08	0.18	0.09	0.07	0.08
16	0.15	0.08	0.08	0.14	0.03	0.08	0.19	0.18	0.12
17	0.12	0.13	0.13	0.01	0.07	0.06	0.04	0.08	0.08
18	0.05	0.12	0.08	0.13	0.09	0.05	0.07	0.05	0.08
19	0.12	0.11	0.05	0.09	0.07	0.12	0.11	0.07	0.09
Average	0.13	0.11	0.09	0.11	0.10	0.12	0.10	0.11	0.11

TABLE 3

Summary of consistency index for correct alternative

Item	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Average
1	0.69	0.76	0.64	0.77	0.85	0.76	0.87	0.73	0.75
2	0.74	0.44	0.62	0.54	0.77	0.57	0.59	0.69	0.62
3	0.76	0.88	0.92	0.85	0.74	0.94	0.88	0.98	0.87
4	0.47	0.78	0.99	0.77	0.99	0.33	0.53	0.42	0.66
5	0.86	1.00	0.82	0.81	0.74	0.80	0.86	0.87	0.85
6	0.92	0.79	0.66	0.59	0.96	0.67	0.90	0.97	0.81
7	0.75	0.87	0.80	0.82	0.48	0.46	0.71	0.76	0.71
8	0.77	0.75	0.73	0.82	0.60	0.54	0.81	0.78	0.73
9	0.97	0.92	0.99	0.58	0.85	0.84	0.99	0.88	0.88
10	0.64	0.72	0.78	0.43	0.99	0.61	0.51	0.68	0.66
11	0.92	0.83	0.82	0.77	0.78	0.61	0.77	0.47	0.75
12	0.49	0.52	0.36	0.30	0.45	0.40	0.46	0.44	0.43
13	0.92	0.95	0.95	0.74	0.88	0.59	0.99	0.94	0.87
14	0.85	0.81	0.71	0.55	0.65	0.81	0.92	0.57	0.74
15	0.82	0.89	0.89	0.94	0.88	0.77	0.69	0.72	0.83
16	0.81	0.91	0.94	0.29	0.90	0.99	0.78	0.89	0.82
17	0.80	0.53	0.56	0.98	0.78	0.78	0.85	0.74	0.75
18	0.96	0.96	0.77	0.39	0.68	0.81	0.81	0.79	0.77
19	0.72	0.73	1.00	0.79	0.83	0.71	0.74	0.84	0.79
Average	0.78	0.77	0.78	0.68	0.78	0.69	0.76	0.74	0.75

TABLE 4

Summary of consistency indices for incorrect alternatives

Item	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Average
1	0.89	0.91	0.88	0.92	0.95	0.88	0.88	0.91	0.90
2	0.92	0.83	0.88	0.85	0.92	0.86	0.88	0.90	0.88
3	0.83	0.91	0.96	0.98	0.90	0.97	0.97	0.94	0.93
4	0.88	0.89	0.91	0.78	0.93	0.76	0.80	0.87	0.85
5	0.72	0.85	0.80	0.95	0.72	0.87	0.89	0.77	0.82
6	0.86	0.89	0.91	0.90	0.88	0.92	0.93	0.95	0.90
7	0.57	0.72	0.92	0.91	0.77	0.84	0.71	0.83	0.78
8	0.86	0.85	0.92	0.84	0.89	0.85	0.92	0.84	0.87
9	0.87	0.93	0.95	0.69	0.93	0.86	0.93	0.95	0.89
10	0.83	0.79	0.90	0.78	0.97	0.88	0.76	0.91	0.85
11	0.72	0.94	0.79	0.76	0.83	0.69	0.87	0.69	0.79
12	0.79	0.83	0.79	0.76	0.73	0.76	0.80	0.74	0.78
13	0.96	0.99	0.93	0.88	0.88	0.92	0.85	0.93	0.92
14	0.88	0.92	0.89	0.83	0.90	0.71	0.93	0.87	0.87
15	0.91	0.90	0.97	0.94	0.92	0.81	0.90	0.92	0.91
16	0.82	0.90	0.90	0.82	0.96	0.90	0.77	0.79	0.86
17	0.87	0.86	0.86	0.99	0.93	0.93	0.95	0.92	0.91
18	0.94	0.86	0.91	0.85	0.91	0.95	0.92	0.95	0.91
19	0.87	0.88	0.95	0.91	0.92	0.87	0.88	0.92	0.90
Average	0.84	0.87	0.90	0.87	0.89	0.86	0.87	0.87	0.87

21

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-01 W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").